2

Automating Supervision of AI Delegates

Jaan Tallinn and Richard Ngo

As the field of machine learning advances, AI systems are becoming more and more useful in a number of domains, in particular due to their increasing ability to generalise beyond their training data. Our focus in this chapter is on understanding the different possibilities for the deployment of highly capable and general systems which we may build in the future. We introduce a framework for the deployment of AI which focuses on two ways for humans to interact with AI systems: delegation and supervision. This framework provides a new lens through which to view both the relationship between humans and AIs, and the relationship between the training and deployment of machine learning systems.

I. AIS AS TOOLS, AGENTS, OR DELEGATES

The last decade has seen dramatic progress in Artificial Intelligence (AI), in particular due to advances in deep learning and reinforcement learning. The increasingly impactful capabilities of our AI systems raise important questions about what future AIs might look like and how we might interact with them. In one sense, AI can be considered a particularly sophisticated type of software. Indeed, the line between AI and other software is very blurry: many software products rely on algorithms which fell under the remit of AI when they were developed, but are no longer typically described as AI.[1] Prominent examples include search engines like Google and image-processing tools like optical character recognition. Thus, when thinking about future AI systems, one natural approach is to picture us interacting with them similarly to how we interact with software programs: as tools which we will use to perform specific tasks, based on predefined affordances, via specially-designed interfaces.

Let us call this the 'tool paradigm' for AI. Although we will undoubtedly continue to develop some AIs which fit under this paradigm, compelling arguments have been made that other AIs will fall outside it – in particular, AIs able to flexibly interact with the real world to perform a wide range of tasks, displaying general rather than narrow intelligence. The example of humans shows that cognitive skills gained in one domain can be useful in a wide range of other domains; it is difficult to argue that the same cannot be true for AIs, especially given the similarities between human brains and deep neural networks. Although no generally intelligent AIs exist today, and some AI researchers are skeptical about the prospects for building them, most expect

[1] M Minsky, 'Thoughts about Artificial Intelligence' in R Kurzweil (ed), *The Age of Intelligent Machines* (1990).

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

it to be possible within this century.[2] However, it does not require particularly confident views on the timelines involved to see value in starting to prepare for the development of artificial general intelligence (AGI) already.

Why won't AGIs fit naturally into the tool paradigm? There are two core reasons: flexibility and autonomy. Tools are built with a certain set of affordances, which allow a user to perform specific tasks with them.[3] For example, software programs provide interfaces for humans to interact with, where different elements of the interface correspond to different functionalities. However, predefined interfaces cannot adequately capture the wide range of tasks that humans are, and AGIs will be, capable of performing. When working with other humans, we solve this problem by using natural language to specify tasks in an expressive and flexible way; we should expect that these and other useful properties will ensure that natural language is a key means of interacting with AGIs. Indeed, AI assistants such as Siri and Alexa are already rapidly moving in this direction.

A second difference between using tools and working with humans: when we ask a human to perform a complex task for us, we don't need to directly specify each possible course of action. Instead, they will often be able to make a range of decisions and react to changing circumstances based on their own judgements. We should expect that, in order to carry out complex tasks like running a company, AGIs will also need to be able to act autonomously over significant periods of time. In such cases, it seems inaccurate to describe them as tools being directly used by humans, because the humans involved may know very little about the specific actions the AGI is taking.

In an extreme case, we can imagine AGIs which possess ingrained goals which they pursue autonomously over arbitrary lengths of time. Let's call this the full autonomy paradigm. Such systems have been discussed extensively by *Nick Bostrom* and *Eliezer Yudkowsky*.[4] *Stuart Russell* argues that they are the logical conclusion of extrapolating the current aims and methods of machine learning.[5] Under this paradigm, AIs would acquire goals during their training process which they then pursue throughout deployment. Those goals might be related to, or influenced by, human preferences and values, but could be pursued without humans necessarily being in control or having veto power.

The prospect of creating another type of entity which independently pursues goals in a similar way to humans raises a host of moral, legal, and safety questions, and may have irreversible effects – because once created, autonomous AIs with broad goals will have incentives to influence human decision-making towards outcomes more favourable to their goals. In particular, concerns have been raised about the difficulty of ensuring that goals acquired by AIs during training are desirable ones from a human perspective. Why might AGIs nevertheless be built with this level of autonomy? The main argument towards this conclusion is that increasing AI autonomy will be a source of competitive economic or political advantage, especially if an AGI race occurs.[6] Once an AI's strategic decision-making abilities exceed those of humans, then the ability to operate independently, without needing to consult humans and wait for their decisions, would

[2] K Grace and others, 'When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 *Journal of Artificial Intelligence Research* 729.

[3] JJ Gibson, 'The Theory of Affordances' in JJ Gibson (ed), *The Ecological Approach to Visual Perception* (1979) 127–137.

[4] N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014); E Yudkowsky, 'Artificial Intelligence as a Positive and Negative Factor in Global Risk' in N Bostrom and MM Cirkovic (eds), *Global Catastrophic Risks* (2008) 184.

[5] S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

[6] S Cave and S ÓhÉigeartaigh, 'An AI Race for Strategic Advantage: Rhetoric and Risks' in J Furman and others (eds), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) 36–40.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

give it a speed advantage over more closely-supervised competitors. This phenomenon has already been observed in high-frequency trading in financial markets – albeit to a limited extent, because trading algorithms can only carry out a narrow range of predefined actions.

Authors who have raised these concerns have primarily suggested that they be solved by developing better techniques for building the *right* autonomous AIs. However, we should not consider it a foregone conclusion that we will build fully autonomous AIs at all. As *Stephen Cave* and *Sean ÓhÉigeartaigh* point out, AI races are driven in part by self-fulfilling narratives – meaning that one way of reducing their likelihood is to provide alternative narratives which don't involve a race to fully autonomous AGI.[7] In this chapter we highlight and explore an alternative which lies between the tool paradigm and the full autonomy paradigm, which we call the supervised delegation paradigm. The core idea is that we should aim to build AIs which can perform tasks and make decisions on our behalf upon request, but which lack persistent goals of their own outside the scope of explicit delegation. Like autonomous AIs, delegate AIs would be able to infer human beliefs and preferences, then flexibly make and implement decisions without human intervention; but like tool AIs, they would lack agency when they have not been deployed by humans. We call systems whose motivations function in this way aligned delegates (as discussed further in the next section).

The concept of delegation has appeared in discussions of agent-based systems going back decades,[8] and is closely related to *Bostrom's* concept of 'genie AI'.[9] Another related concept is the AI assistance paradigm advocated by Stuart Russell, which also focuses on building AIs that pursue human goals rather than their own goals.[10] However, Russell's conception of assistant AIs is much broader in scope than delegate AIs as we have defined them, as we discuss in the next section. More recently, delegation was a core element of *Andrew Critch* and *David Krueger's* ARCHES framework, which highlights the importance of helping multiple humans safely delegate tasks to multiple AIs.[11]

While most of the preceding works were motivated by concern about the difficulty of alignment, they spend relatively little time explaining the specific problems involved in aligning machine learning systems, and how proposed solutions address them. The main contribution of this chapter is to provide a clearer statement of the properties which we should aim to build into AI delegates, the challenges which we should expect, and the techniques which might allow us to overcome them, in the context of modern machine learning (and more specifically deep reinforcement learning). A particular focus is the importance of having large amounts of data which specify desirable behaviour – or, in more poetic terms, the 'unreasonable effectiveness of data'.[12] This is where the supervised aspect of supervised delegation comes in: we argue that, in order for AI delegates to remain trustworthy, it will be necessary to continuously monitor and evaluate their behaviour. We discuss ways in which the difficulties of doing so give rise to a tradeoff between safety and autonomy. We conclude with a discussion of how the goal of alignment can be a focal point for cooperation, rather than competition, between groups involved with AI development.

[7] Ibid.

[8] C Castelfranchi and R Falcone, 'Towards a Theory of Delegation for Agent-Based Systems' (1998) 24(3–4) *Robotics and Autonomous systems* 141.

[9] N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014).

[10] S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

[11] A Critch and D Krueger, 'AI Research Considerations for Human Existential Safety (ARCHES)' (*arXiv*, 30 May 2020) https://arxiv.org/abs/2006.04948v1.

[12] A Halevy, P Norvig, and F Pereira, 'The Unreasonable Effectiveness of Data' (2009) 24(2) *IEEE Intelligent Systems*, 8–12.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

II. ALIGNED DELEGATES

What does it mean for an AI to be aligned with a human? The definition which we will use here comes from *Paul Christiano*: an AI is intent aligned with a human if the AI is trying to do what the human wants it to do.[13] To be clear, this does not require that the AI is correct about what the human wants it to do, nor that it succeeds – both of which will be affected by the difficulty of the task and the AI's capabilities. The concept of intent alignment (henceforth just 'alignment') instead attempts to describe an AI's motivations in a way that's largely separable from its capabilities.

Having said that, the definition still assumes a certain baseline level of capabilities. As defined above, alignment is a property only applicable to AIs with sufficiently sophisticated motivational systems that they can be accurately described as trying to achieve things. It also requires that they possess sufficiently advanced theories of mind to be able to ascribe desires and intentions to humans, and reasonable levels of coherence over time. In practice, because so much of human communication happens via natural language, it will also require sufficient language skills to infer humans' intentions from their speech. Opinions differ on how difficult it is to meet these criteria – some consider it appropriate to take an 'intentional stance' towards a wide range of systems, including simple animals, whereas others have more stringent requirements for ascribing intentionality and theory of mind.[14] We need not take a position on these debates, except to hold that sufficiently advanced AGIs could meet each of these criteria.

Another complication comes from the ambiguity of 'what the human wants'. *Iason Gabriel* argues that 'there are significant differences between AI that aligns with instructions, intentions, revealed preferences, ideal preferences, interests and values'; *Christiano's* definition of alignment doesn't pin down which of these we should focus on.[15] Alignment with the ideal preferences and values of fully-informed versions of ourselves (also known as 'ambitious alignment') has been the primary approach discussed in the context of fully autonomous AI. Even *Russell's* assistant AIs are intended to 'maximise the realisation of human preferences' – where he is specifically referring to preferences that are 'all-encompassing: they cover everything you might care about, arbitrarily far into the future'.[16]

Yet it's not clear whether this level of ambitious alignment is either feasible or desirable. In terms of feasibility, focusing on long timeframes exacerbates many of the problems we discuss in later sections. And in terms of desirability, ambitious alignment implies that a human is no longer an authoritative source for what an AI aligned with that human should aim to do. An AI aligned with a human's revealed preferences, ideal preferences, interests, or values might believe that it understands them better than the human does, which could lead to that AI hiding information from the human or disobeying explicit instructions. Because we are still very far from any holistic theory of human preferences or values, we should be wary of attempts to design AIs which take actions even when their human principals explicitly instruct them not to; let us call this the principle of deference. (Note that the principle is formulated in an asymmetric way – it seems plausible that aligned AIs should sometimes avoid taking actions even when instructed to do so, in particular illegal or unethical actions.)

[13] P Christiano, 'Clarifying "AI Alignment"' (*AI Alignment*, 7 April 2018) https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6.

[14] DC Dennett, 'Précis of the Intentional Stance' (1988) 11 *Behavioral and Brain Stances* 495.

[15] I Gabriel, 'Artificial Intelligence, Values, and Alignment' (2020) 30 *Minds and Machines* 411.

[16] S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

For our purposes, then, we shall define a delegate AI as aligned with a human principal if it tries to do only what that human intends it to do, where the human's intentions are interpreted to be within the scope of tasks the AI has been delegated. What counts as a delegated task depends on what the principal has said to the AI – making natural language an essential element of the supervised delegation paradigm. This contrasts with both the tool paradigm (in which many AIs will not be general enough to understand linguistic instructions) and the full autonomy paradigm (in which language is merely considered one of many information channels which help AIs understand how to pursue their underlying goals).

Defining delegation in terms of speech acts does not, however, imply that all relevant information needs to be stated explicitly. Although early philosophers of language focused heavily on the explicit content of language, more recent approaches have emphasised the importance of a pragmatic focus on speaker intentions and wider context in addition to the literal meanings of the words spoken.[17] From a pragmatic perspective, full linguistic competence includes the ability to understand the (unspoken) implications of a statement, as well as the ability to interpret imprecise or metaphorical claims in the intended way. Aligned AI delegates should use this type of language understanding in order to interpret the 'scope' of tasks in terms of the pragmatics and context of the instructions given, in the same way that humans do when following instructions. An AI with goals which extend outside that scope, or which don't match its instructions, would count as misaligned.

We should be clear that aiming to build aligned delegates with the properties described above will likely involve making some tradeoffs against desirable aspects of autonomy. For example, an aligned delegate would not take actions which are beneficial for its user that are outside the scope of what it has been asked to do; nor will it actively prevent its user from making a series of bad choices. We consider these to be features, though, rather than bugs – they allow us to draw a boundary before reaching full autonomy, with the aim of preventing a gradual slide into building fully autonomous systems before we have a thorough understanding of the costs, benefits, and risks involved. The clearer such boundaries are, the easier it will be to train AIs with corresponding motivations (as we discuss in the next section).

A final (but crucial) consideration is that alignment is a two-place predicate: an AI cannot just be aligned simpliciter, but rather must be aligned with a particular principal – and indeed could be aligned with different principals in different ways. For instance, when AI developers construct an AI, they would like it to obey the instructions given to it by the end user, but only within the scope of whatever terms and conditions have been placed on it. From the perspective of a government, another limitation is desirable: AI should ideally be aligned to their end users only within the scope of legal behaviour. The questions of who AIs should be aligned with, and who should be held responsible for their behaviour, are fundamentally questions of politics and governance rather than technical questions. However, technical advances will affect the landscape of possibilities in important ways. Particularly noteworthy is the effect of AI delegates performing impactful political tasks – such as negotiation, advocacy, or delegation of their own – on behalf of their human principals. The increased complexity of resulting AI governance problems may place stricter requirements on technical approaches to achieving alignment.[18]

[17] K Korta and J Perry, 'Pragmatics' (*The Stanford Encyclopedia of Philosophy*, 21 August 2019) https://plato.stanford.edu/archives/fall2019/entries/pragmatics/.

[18] A Critch and D Krueger, 'AI Research Considerations for Human Existential Safety (ARCHES)' (*arXiv*, 30 May 2020) https://arxiv.org/abs/2006.04948v1.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

III. THE NECESSITY OF HUMAN SUPERVISION

So far we have talked about desirable properties of alignment without any consideration of how to achieve those desiderata. Unfortunately, a growing number of researchers have raised concerns that current machine learning techniques are inadequate for ensuring alignment of AGIs. Research in this area focuses on two core problems. The first is the problem of outer alignment: the difficulty in designing reward functions for reinforcement learning agents which incentivise desirable behaviour while penalising undesirable behaviour.[19] *Victoria Krakovna* et al catalogue many examples of specification gaming in which agents find unexpected ways to score highly even in relatively simple environments, most due to mistakes in how the reward function was specified.[20] As we train agents in increasingly complex and open-ended environments, designing ungameable reward functions will become much more difficult.[21]

One major approach to addressing the problems with explicit reward functions involves generating rewards based on human data – known as reward learning. Early work on reward learning focused on deriving reward functions from human demonstrations – a process known as inverse reinforcement learning.[22] However, this requires humans themselves to be able to perform the task to a reasonable level in order to provide demonstrations. An alternative approach which avoids this limitation involves inferring reward functions from human evaluations of AI behaviour. This approach, known as reward modelling, has been used to train AIs to perform tasks which humans cannot demonstrate well, such as controlling a (simulated) robot body to do a backflip.[23]

In most existing examples of reward learning, reward functions are learned individually for each task of interest – an approach which doesn't scale to systems which generalise to new tasks after deployment. However, a growing body of work on interacting with reinforcement learning agents using natural language has been increasingly successful in training AIs to generalise to novel instructions.[24] This fits well with the vision of aligned delegation described in the previous section, in which specification of tasks for AIs involves two steps: first training AIs to have aligned motivations, and then using verbal instructions to delegate them specific tasks. The hope is that if AIs are rewarded for following a wide range of instructions in a wide range of situations, then they will naturally acquire the motivation to follow human instructions in general, including novel instructions in novel environments.

However, this hope is challenged by a second concern. The problem of inner alignment is that even if we correctly specify the reward function used during training, the resulting policy may not possess the goal described by that reward function. In particular, it may learn to pursue proxy goals which are correlated with reward during most of the training period, but which eventually diverge (either during later stages of training, or during deployment).[25] This possibility is analogous to how humans learned to care directly about food, survival, sex, and so

[19] Alignment problems also exist for AIs trained in other ways, such as self-supervised learning; here I focus on the case of reinforcement learning for the sake of clarity.

[20] V Krakovna and others, 'Specification Gaming: The Flip Side of AI Ingenuity' (*Deep Mind*, 2020) deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

[21] A Ecoffet, J Clune, and J Lehman, 'Open Questions in Creating Safe Open-Ended AI: Tensions between Control and Creativity' in *Artificial Life Conference Proceedings* (2020) 27–35.

[22] AY Ng and SJ Russell, 'Algorithms for Inverse Reinforcement Learning' (2000) in 1 Icml 2.

[23] P Christiano and others, 'Deep Reinforcement Learning from Human Preferences' (*arXiv*, 13 July 2017).

[24] J Luketina and others, 'A Survey of Reinforcement Learning Informed by Natural Language' (*arXiv*,10 June 2019) https://arxiv.org/abs/1906.03926; J Abramson and others, 'Imitating Interactive Intelligence' (*arXiv*, 21 January 2021).

[25] J Koch and others, 'Objective Robustness in Deep Reinforcement Learning'(*arXiv*, 8 June 2021).

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

on – proxies which were strongly correlated with genetic fitness in our ancestral environment, but are much less so today.[26] As an illustration of how inner misalignment might arise in the context of machine learning, consider training a policy to follow human instructions in a virtual environment containing many incapacitating traps. If it is rewarded every time it successfully follows an instruction, then it will learn to avoid becoming incapacitated, as that usually prevents it from completing its assigned task. This is consistent with the policy being aligned, if the policy only cares about surviving the traps as a means to complete its assigned task – in other words, as an instrumental goal. However, if policies which care about survival only as an instrumental goal receive (nearly) the same reward as policies which care about survival for its own sake (as a final goal) then we cannot guarantee that the training process will find one in the former category rather than the latter.

Now, survival is just one example of a proxy goal that might lead to inner misalignment; and it will not be relevant in all training environments. But training environments which are sufficiently complex to give rise to AGI will need to capture at least some of the challenges of the real world – imperfect information, resource limitations, and so on. If solving these challenges is highly correlated with receiving rewards during training, then how can we ensure that policies only learn to care about solving those challenges for instrumental purposes, within the bounds of delegated tasks? The most straightforward approach is to broaden the range of training data used, thereby reducing the correlations between proxy goals and the intended goal. For example, in the environment discussed in the previous paragraph, instructing policies to deliberately walk into traps (and rewarding them for doing so) would make survival less correlated with reward, thereby penalising policies which pursue survival for its own sake.

In practice, though, when talking about training artificial general intelligences to perform a wide range of tasks, we should expect that the training data will encode many incentives which are hard to anticipate in advance. Language models such as GPT-3 are already being used in a wide range of surprising applications, from playing chess (using text interactions only) to generating text adventure games.[27] It will be difficult for AI developers to monitor AI behaviour across many domains, and then design rewards which steer those AIs towards intended behaviour. This difficulty is exacerbated by the fact that modern machine learning techniques are incredibly data-hungry: training agents to perform well in difficult games can take billions of steps. If the default data sources available give rise to inner or outer alignment problems, then the amount of additional supervision required to correct these problems may be infeasible for developers to collect directly. So, how can we obtain enough data to usefully think about alignment failures in a wide range of circumstances, to address the outer and inner alignment problems?

Our suggestion is that this gap in supervision can be filled by end users. Instead of thinking of AI development as a training process followed by a deployment process, we should think of it as an ongoing cycle in which users feed back evaluations which are then used to help align future AIs. In its simplest form, this might involve users identifying inconsistencies or omissions in an AI's statements, or ways in which it misunderstood the user's intentions, or even just occasions when it took actions without having received human instructions. In order to further constrain an AI's autonomy, the AI can also be penalised for behaviour which was desirable, but beyond

[26] E Hubinger and others, 'Risks from Learned Optimization in Advanced Machine Learning Systems' (*arXiv*, 11 June 2019).

[27] S Alexander, 'A Very Unlikely Chess Game' (*Slate Star Codex*, 6 January 2020) https://slatestarcodex.com/2020/01/06/a-very-unlikely-chess-game/;N Walton, 'AI Dungeon: Dragon Model Upgrade' (*Latitude Team*, 14 July 2020) https://aidungeon.medium.com/ai-dungeon-dragon-model-upgrade-7e8ea579abfe.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

the scope of the task it was delegated to perform. This form of evaluation is much easier than trying to evaluate the long-term consequences of an AI's actions; yet it still pushes back against the underlying pressure towards convergent instrumental goals and greater autonomy that we described above.

Of course, user data is already collected by many different groups for many different purposes. Prominent examples include scraping text from Reddit, or videos from YouTube, in order to train large self-supervised machine learning models. However, these corpora contain many examples of behaviour we wouldn't like AIs to imitate – as seen in GPT-3's regurgitation of stereotypes and biases found in its training data.[28] In other cases, evaluations are inferred from user behaviour: likes on a social media post, or clicks on a search result, can be interpreted as positive feedback. Yet these types of metrics already have serious limitations: there are many motivations driving user engagement, not all of which should be interpreted as positive feedback. As interactions with AI become much more freeform and wide-ranging, inferred correlations will become even less reliable, compared with asking users to evaluate AI alignment directly. So even if users only perform explicit evaluations of a small fraction of AI behaviour, this could provide much more information about their alignment than any other sources of data currently available. And, unlike other data sources, user evaluations could flexibly match the distributions of tasks on which AIs are actually deployed in the real world, and respond to new AI behaviour very quickly.[29]

IV. BEYOND HUMAN SUPERVISION

Unfortunately, there are a number of reasons to expect that even widespread use of human evaluation will not be sufficient for reliable supervision in the long term. The core problem is that the more sophisticated an AI's capabilities are, the harder it is to identify whether it is behaving as intended or not. In some narrow domains like chess and Go, experts already struggle to evaluate the quality of AI moves, and to tell the difference between blunders and strokes of brilliance. The much greater complexity of the real world will make it even harder to identify all the consequences of decisions made by AIs, especially in domains where they make decisions far faster and generate much more data than humans can keep up with.

Particularly worrying is the possibility of AIs developing deceptive behaviour with the aim of manipulating humans into giving better feedback. The most notable example of this came from reward modelling experiments in which a human rewarded an AI for grasping a ball with a robotic claw.[30] Instead of completing the intended task, the AI learned to move the claw into a position between the camera and the ball, thus appearing to grasp the ball without the difficulty of actually doing so. As AIs develop a better understanding of human psychology and the real-world context in which they're being trained, manipulative strategies like this could become much more complex and much harder to detect. They would also not necessarily be limited to affecting observations sent directly to humans, but might also attempt to modify their reward signal using any other mechanisms they can gain access to.

[28] TB Brown and others, 'Language Models Are Few-Shot Learners' (*arXiv*, 22 July 2020) https://arxiv.org/abs/2005.14165?source=techstories.org.

[29] This does assume a high level of buy-in from potential users, which may be difficult to obtain given privacy concerns. We hope that the project of alignment can be presented in a way that allows widespread collaboration – as discussed further in the final section of this chapter.

[30] P Christiano and others, 'Deep Reinforcement Learning from Human Preferences' (*arXiv*, 13 July 2017) https://arxiv.org/abs/1706.03741.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

The possibility of manipulation is not an incidental problem, but rather a core difficulty baked into the use of reinforcement learning in the real world. As AI pioneer *Stuart Russell* puts it:

> The formal model of reinforcement learning assumes that the reward signal reaches the agent from outside the environment; but [in fact] the human and robot are part of the same environment, and the robot can maximize its reward by modifying the human to provide a maximal reward signal at all times. . . . [This] indicates a fundamental flaw in the standard formulation of RL.[31]

In other words, AIs are trained to score well on their reward functions by taking actions to influence the environment around them, and human supervisors are a part of their environment which has a significant effect on the reward they receive, so we should expect that by default AIs will learn to influence their human supervisors. This can be mitigated if supervisors heavily penalise attempted manipulation when they spot it – but this still leaves an incentive for manipulation which can't be easily detected. As AIs come to surpass human abilities on complex real-world tasks, preventing them from learning manipulative strategies will become increasingly difficult – especially if AI capabilities advance rapidly, so that users and researchers have little time to notice and respond to the problem.

How might we prevent this, if detecting manipulation or other undesirable behaviour eventually requires a higher quality and quantity of evaluation data than unaided humans can produce? The main mechanisms which have been proposed for expanding the quality/quantity frontier of supervision involve relying on AI systems themselves to help us supervise other AIs. When considering this possibility, we can reuse two of the categories discussed in the first section: we can either have AI-based supervision tools, or else we can delegate the process of supervision to another AI (which we shall call recursive supervision, as it involves an AI delegate supervising another AI delegate, which might then supervise another AI delegate, which…).

One example of an AI-based supervision tool is a reward model which learns to imitate human evaluations of AI behaviour. Reinforcement learning agents whose training is too lengthy for humans to supervise directly (e.g. involving billions of steps) can then be supervised primarily by reward models instead. Early work on reward models demonstrated a surprising level of data efficiency: reward models can greatly amplify a given amount of human feedback.[32] However, the results of these experiments also highlighted the importance of continual feedback – when humans stopped providing new data, agents eventually found undesirable behaviours which nevertheless made the reward models output high scores.[33] So reward models are likely to rely on humans continually evaluating AI behaviour as it expands into new domains.

Another important category of supervision tool is interpretability tools, which aim to explain the mechanisms by which a system decides how to act. Although deep neural networks are generally very opaque to mechanistic explanation, there has been significant progress over the last few years in identifying how groups of artificial neurons (and even individual neurons) contribute to the overall output.[34] One long-term goal of this research is to ensure that AIs will honestly explain the reasoning that led to their actions and their future intentions. This would help address the inner alignment problems described above, because agents could be penalised

[31] S Russell, 'Provably Beneficial Artificial Intelligence' in A de Grey and others (eds), *Exponential Life, The Next Step* (2017).

[32] P Christiano and others, 'Deep Reinforcement Learning from Human Preferences' (*arXiv*, 13 July 2017) https://arxiv.org/abs/1706.03741.

[33] B Ibarz and others, 'Reward Learning from Human Preferences and Demonstrations in Atari' (*arXiv*, 15 November 2018) https://arxiv.org/abs/1811.06521.

[34] N Cammarata and others, 'Thread: Circuits' (*Distill*, 10 March 2020) https://distill.pub/2020/circuits/.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

for acting according to undesirable motivations even when their behaviour is indistinguishable from the intended behaviour. However, existing techniques are still far from being able to identify deceptiveness (or other comparably abstract traits) in sophisticated models.

Recursive supervision is currently also in a speculative position, but some promising strategies have been identified. A notable example is *Geoffrey Irving*, *Paul Christiano*, and *Dario Amodei's* Debate technique, in which two AIs are trained to give arguments for opposing conclusions, with a human judging which arguments are more persuasive.[35] Because the rewards given for winning the debate are zero-sum, the resulting competitive dynamic should in theory lead each AI to converge towards presenting compelling arguments which are hard to rebut – analogous to how AIs trained via self-play on zero-sum games converge to winning strategies. However, two bottlenecks exist: the ease with which debaters can identify flaws in deceptive arguments, and the accuracy with which humans can judge allegations of deception. Several strategies have been proposed to make judging easier – for example, incorporating cross-examination of debaters, or real-world tests of claims made during the debate – but much remains to be done in fleshing out and testing Debate and other forms of recursive supervision.

To some extent, recursive supervision will also arise naturally when multiple AIs are deployed in real-world scenarios. For example, if one self-driving car is driving erratically, then it's useful for others around it to notice and track that. Similarly, if one trading AI is taking extreme positions that move the market considerably, then it's natural for other trading AIs to try to identify what's happening and why. This information could just be used to flag the culprit for further investigation – but it could also be used as a supervision signal for further training, if shared with the relevant AI developers. In the next section we discuss the incentives which might lead different groups to share such information, or to cooperate in other ways.

V. AI SUPERVISION AS A COOPERATIVE ENDEAVOUR

We started this chapter by discussing some of the competitive dynamics which might be involved in AGI development. However, there is reason to hope that the process of increasing AI alignment is much more cooperative than the process of increasing AI capabilities. This is because misalignment could give rise to major negative externalities, especially if misaligned AIs are able to accumulate significant political, economic, or technological power (all of which are convergent instrumental goals). While we might think that it will be easy to 'pull the plug' on misbehaviour, this intuition fails to account for strategies which highly capable AIs might use to prevent us from doing so – especially those available to them after they have already amassed significant power. Indeed, the history of corporations showcases a range of ways that 'agents' with large-scale goals and economic power can evade oversight from the rest of society. And AIs might have much greater advantages than corporations currently do in avoiding accountability – for example, if they operate at speeds too fast for humans to monitor. One particularly stark example of how rapidly AI behaviour can spiral out of control was the 2010 Flash Crash, in which high-frequency trading algorithms got into a positive feedback loop and sent prices crashing within a matter of minutes.[36] Although the algorithms involved were relatively simple by the standards of modern machine learning (making this an example of accidental failure rather than misalignment),

[35] G Irving, P Christiano, and D Amodei, 'AI Safety via Debate' (*arXiv*, 22 October 2018) https://arxiv.org/abs/1805.00899.

[36] US Securities & Exchange Commission and US Commodity Futures Trading Commission, 'Findings Regarding the Market Events of May 6, 2010. Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues' (*US Securities and Exchange Commission*, 30 September 2010) www.sec.gov/news/studies/2010/marketevents-report.pdf.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

AIs sophisticated enough to reason about the wider world will be able to deliberately implement fraudulent or risky behaviour at increasingly bewildering scales.

Preventing them from doing so is in the interests of humanity as a whole; but what might large-scale cooperation to improve alignment actually look like? One possibility involves the sharing of important data – in particular data which is mainly helpful for increasing AI's alignment rather than their capabilities. It is somewhat difficult to think about how this would work for current systems, as they don't have the capabilities identified as prerequisites for being aligned in section 2. But as one intuition pump for how sharing data can differentially promote safety over other capabilities, consider the case of self-driving cars. The data collected by those cars during deployment is one of the main sources of competitive advantage for the companies racing towards autonomous driving, making them rush to get cars on the road. Yet, of that data, only a tiny fraction consists of cases where humans are forced to manually override the car steering, or where the car crashes. So while it would be unreasonably anticompetitive to force self-driving car companies to share all their data, it seems likely that there is some level of disclosure which contributes to preventing serious failures much more than to erasing other competitive advantages. This data could be presented in the form of safety benchmarks, simple prototypes of which include DeepMind's AI Safety Gridworlds and the Partnership on AI's SafeLife environment.[37]

The example of self-driving cars also highlights another factor which could make an important contribution to alignment research: increased cooperation amongst researchers thinking about potential risks from AI. There is currently a notable divide between researchers primarily concerned about near-term risks and those primarily concerned about long-term risks.[38] Currently, the former tend to focus on supervising the activity of existing systems, whereas the latter prioritise automating the supervision of future systems advanced enough to be qualitatively different from existing systems. But in order to understand how to supervise future AI systems, it will be valuable to have access not only to technical research on scalable supervision techniques, but also to hands-on experience of how supervision of AIs works in real-world contexts and the best practices identified so far. So, as technologies like self-driving cars become increasingly important, we hope that the lessons learned from their deployment can help inform work on long-term risks via collaboration between the two camps.

A third type of cooperation to further alignment involves slowing down capabilities research to allow more time for alignment research to occur. This would require either significant trust between the different parties involved, or else strong enforcement mechanisms.[39] However, cooperation can be made easier in a number of ways. For example, corporations can make themselves more trustworthy via legal commitments such as windfall clauses.[40] A version of this has already been implemented in OpenAI's capped-profit structure, along with other innovative legal mechanisms – most notably the clause in OpenAI's charter which commits to assisting rather than competing with other projects, if they meet certain conditions.[41]

[37] J Leike and others, 'AI Safety Gridworlds' (*arXiv*, 28 November 2017) https://arxiv.org/abs/1711.09883; CL Wainwright and P Eckersley, 'Safelife 1.0: Exploring Side Effects in Complex Environments' (*arXiv*, 26 February 2021) https://arxiv.org/abs/1912.01217.

[38] S Cave and S ÓhÉigeartaigh, 'Bridging Near-and Long-Term Concerns about AI' (2019) 1 *Nature Machine Intelligence* 5–6.

[39] A Dafoe, 'AI Governance: A Research Agenda' in *Governance of AI Program*, Future of Humanity Institute, 2017 www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf, 1–53.

[40] C O'Keefe and others, 'The Windfall Clause: Distributing the Benefits of AI for the Common Good' in J Furman and others, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) 327–331.

[41] OpenAI, 'OpenAI Charter' (9 April 2018) https://openai.com/charter/; OpenAI, 'OpenAI LP' (11 March 2019) https://openai.com/blog/openai-lp/.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press

We are aware that we have skipped over many of the details required to practically implement large-scale cooperation to increase AI alignment – some of which might not be pinned down for decades to come. Yet we consider it important to raise and discuss these ideas relatively early, because they require a few key actors (such as technology companies and the AI researchers working for them) to take actions whose benefits will accrue to a much wider population – potentially all of humanity. Thus, we should expect that the default incentives at play will lead to underinvestment in alignment research.[42] The earlier we can understand the risks involved, and the possible ways to avoid them, the easier it will be to build a consensus about the best path forward which is strong enough to overcome whatever self-interested or competitive incentives push in other directions. So despite the inherent difficulty of making arguments about how technological progress will play out, further research into these ideas seems vital for reducing the risk of humanity being left unprepared for the development of AGI.

[42] S Armstrong, N Bostrom, and C Shulman, 'Racing to the Precipice: A Model of Artificial Intelligence Development' (2016) 31 *AI & Society* 201.

https://doi.org/10.1017/9781009207898.004 Published online by Cambridge University Press